

## Education

- 2023 – 2028 ■ **Ph.D. in Computer Science, UT Austin.** Advisor: Aditya Akella and Venkat Arun. Research interest: Computer networks, systems for machine learning.
- 2019 – 2023 ■ **(Summa Cum Laude) B.S. in Computer Science, Turing Class, Peking University.** GPA: 3.869/4 (**rank 2/200**). Advisor: Tong Yang. Research interest: Databases and theoretical computer science. Thesis: “Chainedfilter: Combining membership filters by chain rule” (*Outstanding Undergraduate Thesis* (3%)).

## Research Experience

- 2023 ■ Graduate Research Assistant @ UT Austin, advised by Prof. Daehyeok Kim and Aditya Akella.  
**Topic: Homomorphic and Lossless Compression that Merges Worker-level Compression with In-network Aggregation for Accelerating Distributed DNN Training.**  
 As deep neural networks (DNNs) grow in complexity and size, the resultant increase in communication overhead during distributed training has become a significant bottleneck, challenging the scalability of distributed training systems. Existing solutions, while aiming to mitigate this bottleneck through worker-level compression and in-network aggregation, fall short due to their inability to efficiently reconcile the trade-offs between compression effectiveness and computational overhead, hindering overall performance and scalability. During this period, *I propose a novel compression algorithm that effectively merges worker-level compression with in-network aggregation.* The algorithm is both homomorphic, allowing for efficient in-network aggregation without CPU/GPU processing, and lossless, ensuring no compromise on training accuracy. *Theoretically optimal in compression ratio and computational complexity, our approach is empirically validated across diverse DNN models such as NCF, LSTM, VGG19, and BERT-base, showing up to a  $6.33\times$  improvement in aggregation throughput and a  $3.74\times$  increase in per-iteration training speed.*
- Research Assistant @ UT Austin, advised by Prof. Aditya Akella.  
**Topic: Compact In-network Membership Query Data Structures.**  
 Given a collection of key-value pairs, where each value is either 0 or 1, *membership* aims to accurately identify these pairs, allowing for Type I (false positive) errors. Membership plays an important role in networking, databases, and security. Although have been studied for over fifty years, the space lower bound for general membership problems was unknown. During this period, I provides the space lower bound and gives a surprising theoretical result of membership. Let's first consider a scenario with  $n$  mappings to one and  $\lambda n$  mappings to zero. We can express the space lower bound for this problem as  $n f(\epsilon, \lambda) + o(n)$ . It's easy to see that the problem can be decomposed into two sub-problems: First, storing the  $n$  one mapping and the  $\lambda n$  zero mappings with a false positive rate of  $\epsilon' \in [\epsilon, 1]$ , and second, storing the  $n$  one mappings and the  $\epsilon' \lambda n$  remaining false positive zero mappings with a false positive rate of  $\epsilon/\epsilon'$ . This leads to the inequality  $f(\epsilon, \lambda) \leq f(\epsilon', \lambda) + f(\epsilon/\epsilon', \epsilon' \lambda)$ . The main contribution of ChainedFilter is that the proof of

$$f(\epsilon, \lambda) = f(\epsilon', \lambda) + f(\epsilon/\epsilon', \epsilon' \lambda),$$

indicating that the decomposition process described above involves zero information loss. *This discovery allows me to establish for the first time a complete space lower bound for general membership problems, i.e.,  $f(\epsilon, \lambda) = f(0, \lambda) - f(0, \epsilon \lambda)$ , after the expressions  $f(0, \lambda)$  and  $f(\epsilon, +\infty)$  were known in 1978.* This theory shows that by combining two sub-algorithms, an effective membership algorithm can be developed. Both theoretical and experimental results show that this technique significantly improves the performance of many fundamental applications, including static dictionaries, lossless data compression, cuckoo hashing, learned filters, and LSM trees in RocksDB. This work is accepted by SIGMOD'24.

## Research Experience (continued)

- 2022
- Research Assistant @ Harvard University, advised by Prof. Minlan Yu. During this period, I worked on a switch architecture project to introduce 2d-mesh structure into programmable switches. I write verilog code of a NetCache instance and verify its correctness with simulator.
  - Research Assistant @ Peking University, advised by Prof. Tong Yang.  
**Topic: Fast and Accurate Frequency Estimation for Network Measurement.**  
Recording the frequency of items in highly skewed data streams is a fundamental and hot problem in recent years. The literature demonstrates that sketch is the most promising solution. The typical metrics to measure a sketch are accuracy and speed, but existing sketches make only trade-offs between the two dimensions. During this period, I propose a new sketch framework called Stingy sketch with two key techniques: Bit-pinching Counter Tree (BCTree) and Prophet Queue (PQueue) which optimizes both the accuracy and speed. The key idea of BCTree is to split a large fixed-size counter into many small nodes of a tree structure, and to use a precise encoding to perform carry-in operations with low processing overhead. The key idea of PQueue is to use pipelined prefetch technique to make most memory accesses happen in L2 cache without losing precision. Importantly, the two techniques are cooperative so that Stingy sketch can improve accuracy and speed simultaneously. Extensive experimental results show that Stingy sketch is up to 50% more accurate than the SOTA of accuracy-oriented sketches and is up to 33% faster than the SOTA of speed-oriented sketches. This work is accepted by VLDB'22.
  - Research Assistant @ **Peking University**, advised by Prof. **Tong Yang**. I worked on probabilistic structures for data streams. To be more specific, I focus on fundamental algorithm designs including (2) a novel relational query index for Neo4j database (accepted by ICDE'23), (3) a fast index encoder for range query (accepted by ICDE'23) and (4) quantile estimation algorithms (in submission to ICDE'24).
- 2024
- Existing research directions and projects.
    - Scheduling. Real-time scheduler for robotic operating system (ROS).
    - Programmable hardware. Compact and fast hash table for programmable data plane.
    - DB query optimization. Reducing latency in retrieval augmented generation (RAG) databases and accelerating LLM inference; Query optimization for B+ tree.
    - Learned OS. Learned cache for operating systems.

## Honors and Awards

### Undergraduate

- Scholarships
- Zhongying Moral Education Scholarship (CNY 16,000), 2020-2023.
  - SenseTime Scholarship (30 in China, CNY 20,000), 2022.
  - Arawana Scholarship (CNY 12,000), 2022.
  - National Scholarship (1%, CNY 16,000), 2021.
  - National Scholarship (1%, CNY 16,000), 2020.
- Honors
- Outstanding Graduate, Beijing (5%), 2023.
  - Outstanding Graduate, Peking University, 2023.
  - Weiming Scholar (CNY 10,000), 2023.
  - Merit Student (1/26), 2023.
  - Outstanding Undergraduate Thesis, Peking University (3%), 2023.




## Honors and Awards (continued)

---

- Awards
- **Merit Student Pacesetter (1/26), 2021.**
  - **Merit Student (1/30), 2020.**
  - **Outstanding Undergraduate Research Project (4,000 CNY), 2023.**
  - **Academic Innovation Award (1%), 2022.**
  - **First Prize of Challenge Cup Research Competition (1/11, 3,000 CNY), 2022.**
  - **First Prize of Huawei Cup Innovation Competition (1/16, 4,000 CNY), 2022.**

## Publications







---

- 1 H. Li, Q. Chen, Y. Zhang, T. Yang, and B. Cui, “Stingy sketch: A sketch framework for accurate and fast frequency estimation,” *Proc. VLDB Endow.*, vol. 15, no. 7, pp. 1426–1438, Mar. 2022 (**Top conference in databases**), ISSN: 2150-8097.  DOI: 10.14778/3523210.3523220.
- 2 H. Li, L. Wang, Q. Chen, *et al.*, “Chainedfilter: Combining membership filters by chain rule,” *Proc. ACM Manag. Data*, vol. 1, no. 4, Dec. 2023 (**Top conference in databases**).  DOI: 10.1145/3626721.
- 3 J. Guo, Q. Lyu, Y. Wu, *et al.*, “Qsketch: Per-key quantile estimation centered at one point,” in *submission to ICDE’24*.
- 4 H. Li, Y. Xu, J. Chen, *et al.*, “Accelerating distributed deep learning using lossless homomorphic compression,” in *submission to ICML’24*.  URL: <https://arxiv.org/abs/2402.07529>.
- 5 X. Li, Z. Fan, H. Li, *et al.*, “Steadysketch: Finding steady flows in data streams,” in *2023 IEEE/ACM 31st International Symposium on Quality of Service (IWQoS)*, IEEE, 2023, pp. 01–09.
- 6 R. Qiu, Y. Ming, Y. Hong, H. Li, and T. Yang, “Wind-bell index: Towards ultra-fast edge query for graph databases,” in *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, IEEE, 2023, pp. 2090–2098.
- 7 Z. Wang, Z. Zhong, J. Guo, *et al.*, “Rencoder: A space-time efficient range filter with local encoder,” in *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, IEEE, 2023, pp. 2036–2049.

## Miscellaneous Experience

---

### Undergraduate

- |             |  |
|-------------|--|
| 2020 – 2023 |  Turing Class Monitor.                                    |
| 2022        |  Chair of Turing Student Research Forum.                  |
| 2021        |  Teaching Assistant for Introduction to Computer Systems. |
|             |  Freshmen Counselor.                                      |
|             |  Leader of the College Debating Team.                     |
|             |  Minister of the Zhongying Public Welfare Association.    |